Lab Project 1

Course : Introduction to Probability and Statistics, Math 113 Section 3234
Instructor: Abhijit Champanerkar
Date: Oct 3rd 2012

Empirical rule

The **Empirical Rule** or the **68-95-99.7 Rule** states that for if a frequency distribution of a set of sample data is **normally distributed** then

- Approximately 68% of the data falls within 1 standard deviation of the mean i.e. within $\overline{x} \pm s$.
- Approximately 95% of the data falls within 2 standard deviations of the mean i.e. within $\overline{x} \pm 2s$.
- Approximately 99.7% of the data falls within 3 standard deviations of the mean i.e. within $\overline{x} \pm 3s$.

Dataset "faithful"

Lets test the data set "the length of time of eruptions of the Old Faithful Geyser in Yellowstone" to see if it satisfies the Empirical Rule. In R, load the data:

```
> data(faithful)
> help(faithful)  %% Shows info on the data
> attach(faithful)  %% Loads the 'names' in faithful into R
> names(faithful)
```

Lets look at the eruption times now in variable `eruptions`.

```
> length(eruptions)      %% How much data
> hist(eruptions,20)
```

We want to see what percentage of the data is within one, two and three standard deviations. Lets compute the mean and the standard deviation and save the numbers:

```
> mean(eruptions)
> sd(eruptions)
```

Let us save the mean and standard deviation in the variables `emean` and textttesd respecitvely.

```
> emean=mean(eruptions)
> esd=sd(eruptions)
```

To find the number of observations within ± 1 standard deviation of the mean we use the **R** command `sum`:

```
> sum(erupstions>emean-esd & eruptions < emean+esd)
```

We can get the percentage by dividing by the number of observations which is given by `length` function:

```
> sum(erupstions>emean-esd & eruptions < emean+esd)/length(eruptions)
```

The answer is 55.1%. This means that the data set eruptions DOES NOT satisfy the Empirical rule, which in turn means that the data set is NOT normally distributed. This can also be seen from the histogram.

Questions

1. For the above data set, what percentage of data falls in the range of $\overline{x} \pm 1.25s$?
2. For the above data set, what percentage of data falls in the range of $\overline{x} \pm 2s$, $\overline{x} \pm 3s$?
3. How does it compare with the Empirical Rule ?
4. Find percentages of data which falls in the range $\overline{x} \pm s$, $\overline{x} \pm 1.25s$, $\overline{x} \pm 2s$ and $\overline{x} \pm 3s$ for the data set called "AirPassengers" and "LakeHuron". Compare your results with Empirical Rule.

To hand in:

1. 3 Histograms for the data sets Faithful, AirPassengers and LakeHuron with your name, mean and standard deviation typed out.
2. Type out the answers to Questions 1, 2 and 3 for the data sets Faithful, AirPassengers and LakeHuron in a word processor.
